Probability and Sampling Error

Understanding the Bridge Between Data and Inference

Movement Science Statistics Course | Chapter 8

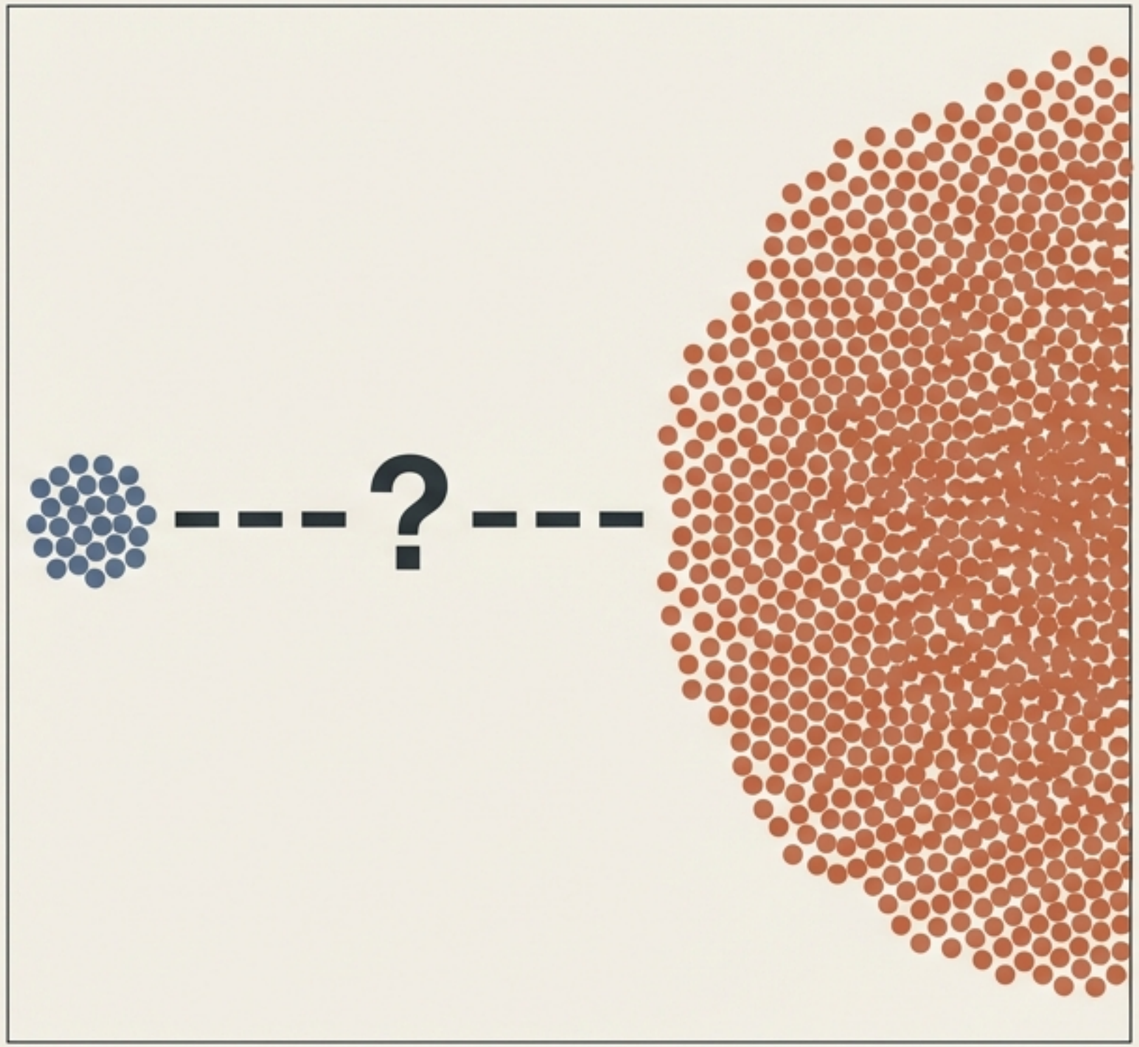
NotebookLM

The Fundamental Challenge of Research

In Movement Science, we seek truths about entire populations—for example, all collegiate athletes or every patient with Parkinson's disease. However, it is impossible to measure everyone. We only ever possess data from a small subset, known as a sample.

This constraint creates a gap between what we have (sample data) and what we want (population truths). How can we be sure our small slice of reality accurately reflects the whole pie?

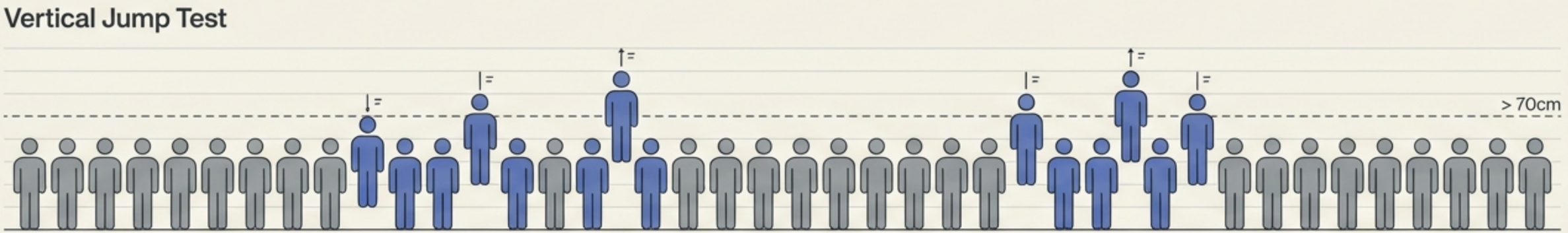
Key Definition: Statistical Inference is the process of drawing conclusions about populations based on limited sample data.



NotebookLM

Probability: The Language of Uncertainty

Probability is the mathematical tool we use to bridge the gap between sample and population. In research, we view probability as **the long-run relative frequency** of an event. It asks: “If we could repeat this study infinitely many times, how often would this result occur?”



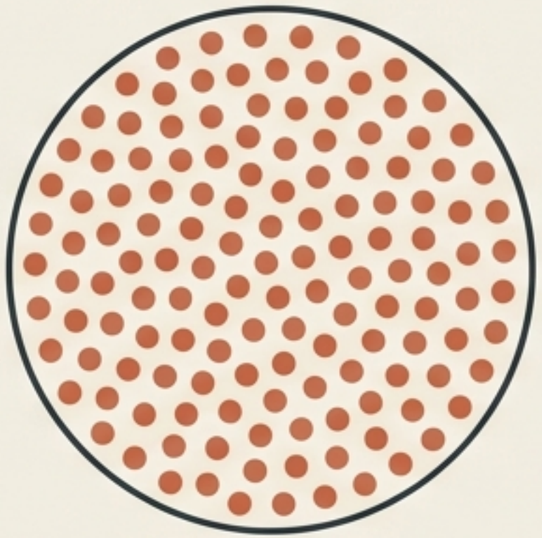
If 25 out of 80 players jump > 70cm, the empirical probability is 25 ÷ 80 = 0.31 (31%).

We use this logic to quantify uncertainty. We don’t just ask if an event happened; we calculate the likelihood of it happening again due to chance.

NotebookLM

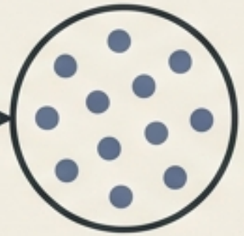
The Players: Population vs. Sample

The Population



- The entire group of interest (e.g., all professional soccer players).
- Described by **Parameters**.
- Notation: Mean (μ) and Standard Deviation (σ).
- Note: These are typically unknown truths.

The Sample



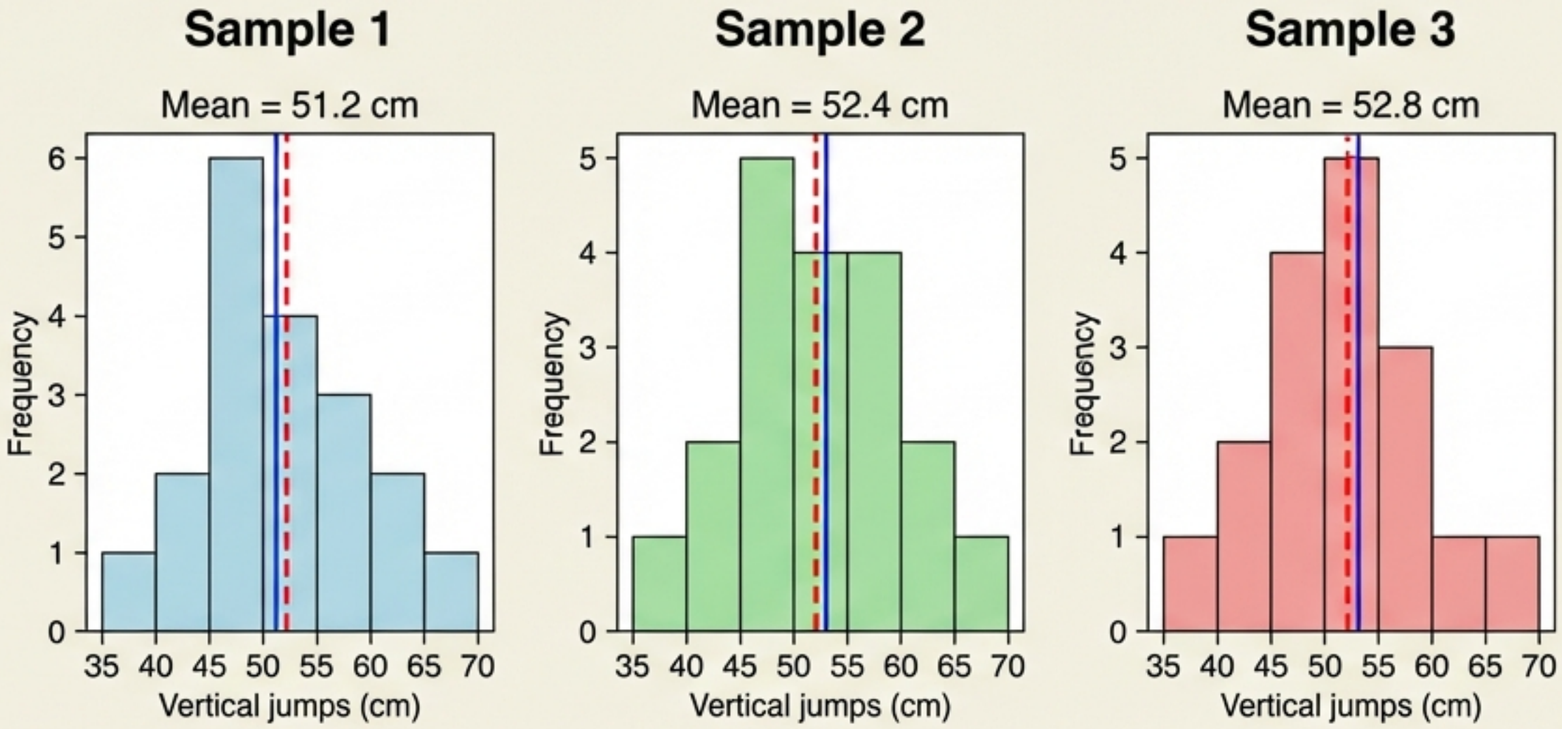
- The subset we actually measure (e.g., the 30 players in our lab).
- Described by **Statistics**.
- Notation: Mean ($\bar{x}$) and Standard Deviation (s).
- Note: These are estimates used to guess the parameters.

NotebookLM

Sampling Error is Not a Mistake

If you take three different groups of athletes and measure their vertical jump, you will get three different averages. This variability is called **Sampling Error**.

It is not a calculation error or a flaw in your study design; it is the unavoidable consequence of looking at a part rather than the whole.



Notice that Sample 1 yields a mean of 51.2 cm, while Sample 2 yields 52.4 cm. Both come from the same population (μ = 52 cm), yet neither is exactly correct. This “wobble” around the truth is sampling error.

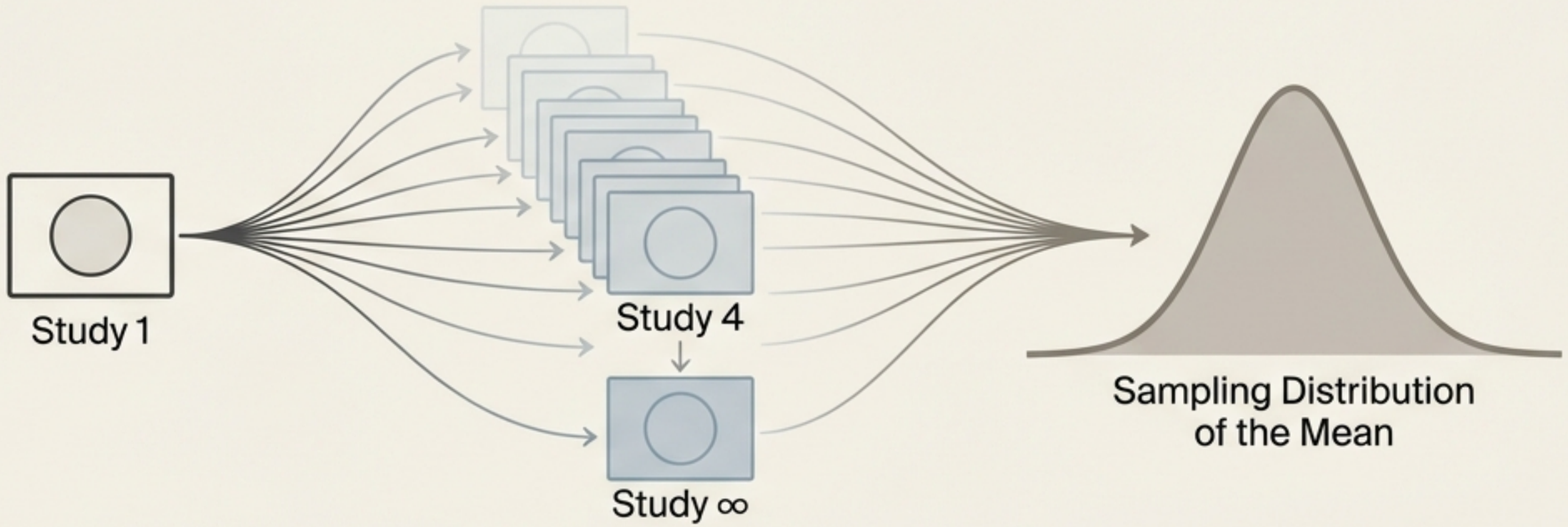
NotebookLM

The Theoretical Bridge: The Sampling Distribution

Imagine if, instead of running your study once, you ran it infinite times. You would have thousands of sample means. If you plotted all those means on a graph, you would create a **Sampling Distribution**.

We never see this distribution in real life (we only do one study), but statistical theory tells us exactly what it looks like.

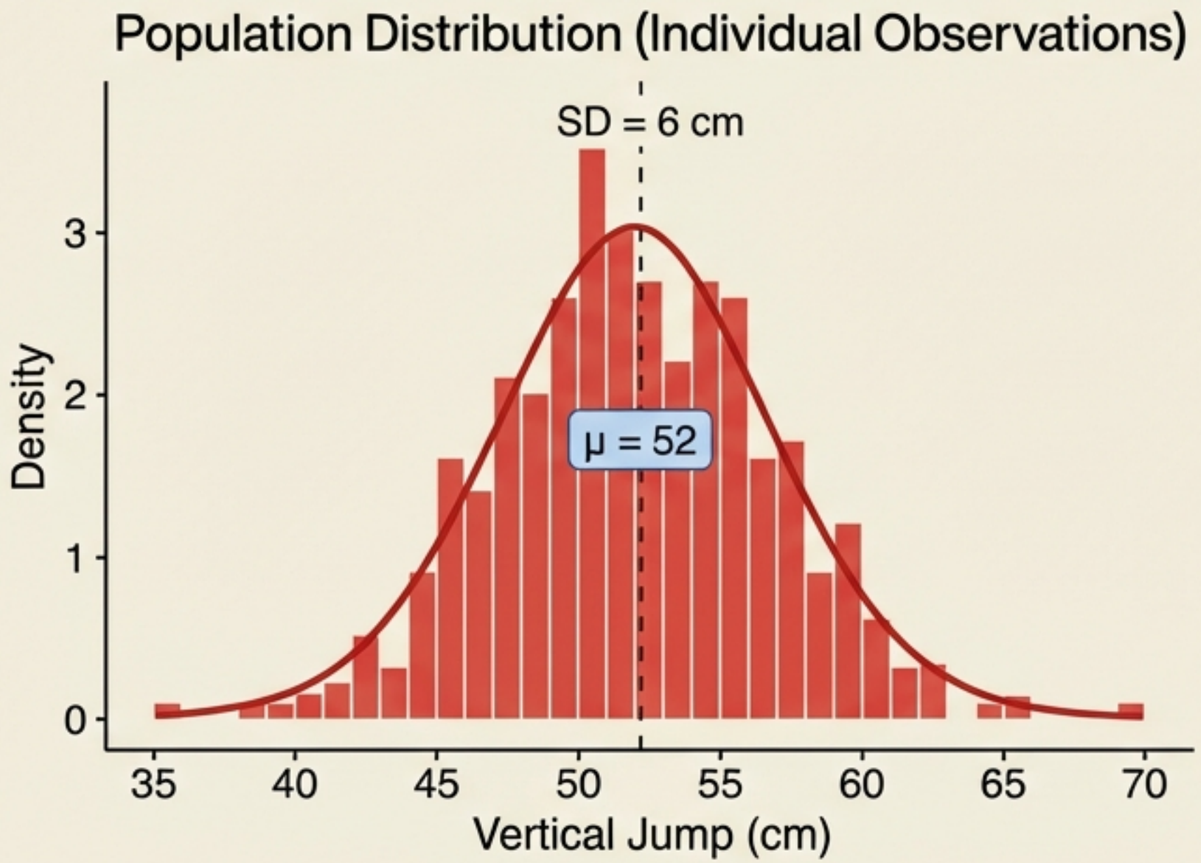
This theoretical distribution bridges the gap: it predicts exactly how much our sample statistics should vary by chance, allowing us to judge if our specific result is “normal” or “surprising.”

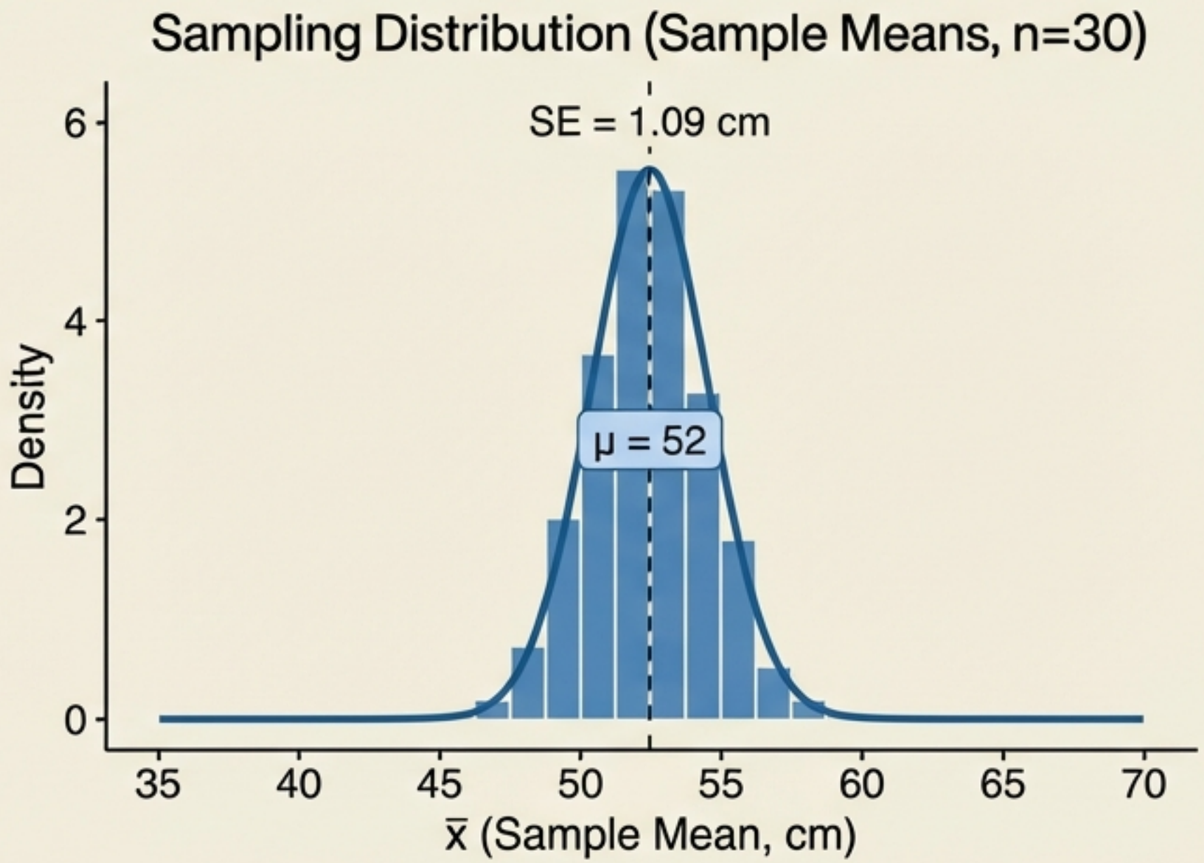


NotebookLM

Visualizing the Shift: Individuals vs. Means

There is a critical difference between the population distribution and the sampling distribution. Individual athletes vary widely. However, the *average* of a group of 30 athletes will cluster tightly around the true population mean. Means are less variable than individual observations.



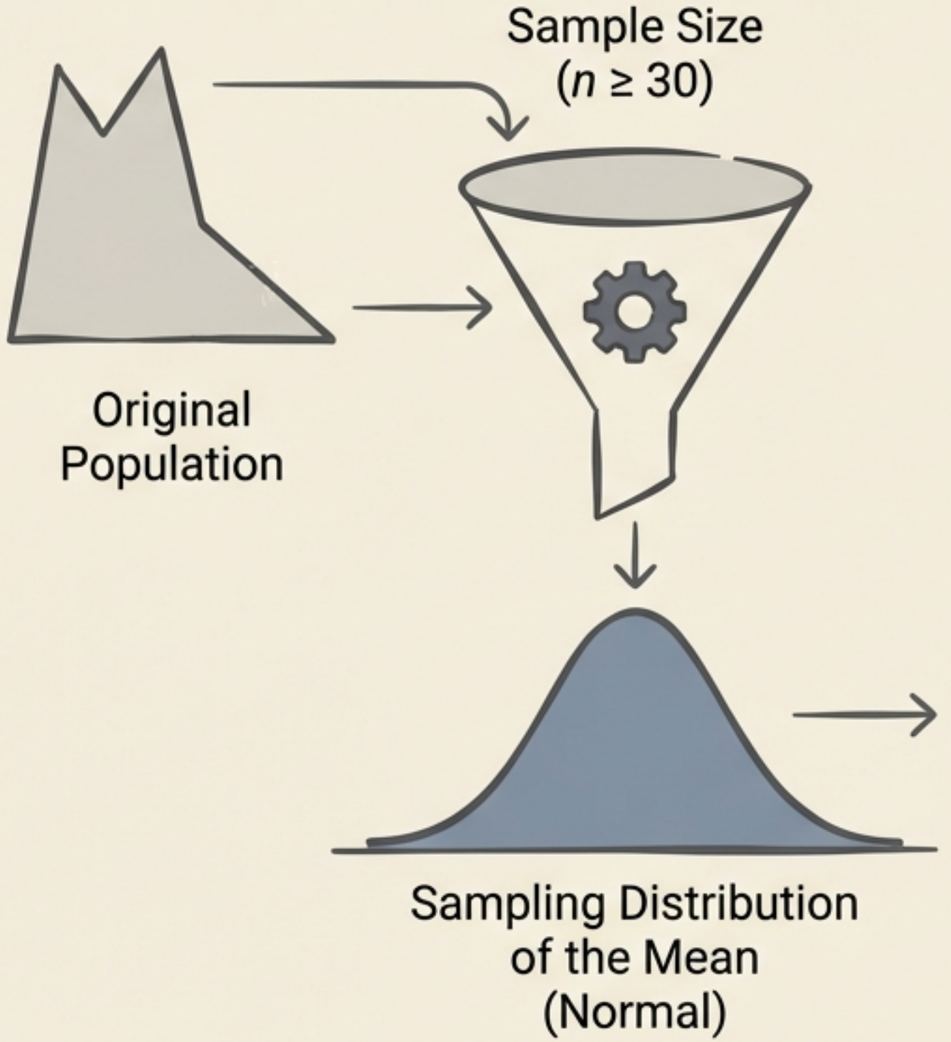


NotebookLM

The Central Limit Theorem (CLT)

The Central Limit Theorem provides a powerful guarantee: as your sample size increases, the sampling distribution of the mean becomes a **Normal Distribution** (bell-shaped), even if the original population is not normal. This "magic" generally kicks in when the sample size is adequate (typically $n \geq 30$).

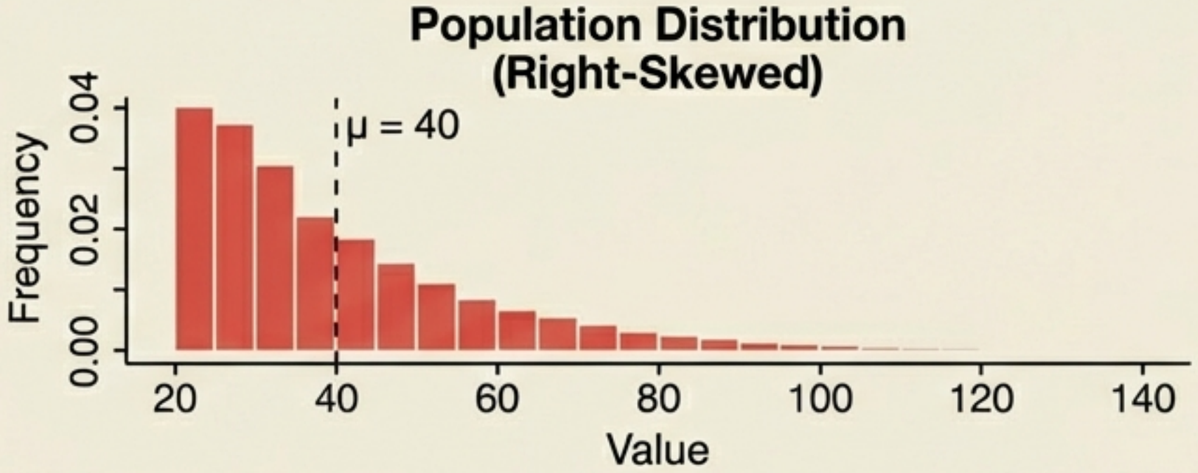
Implication: This is why we can use normal distribution theory to make inferences about means (like t-tests and ANOVAs) even when measuring variables that might be skewed, provided we have enough participants.

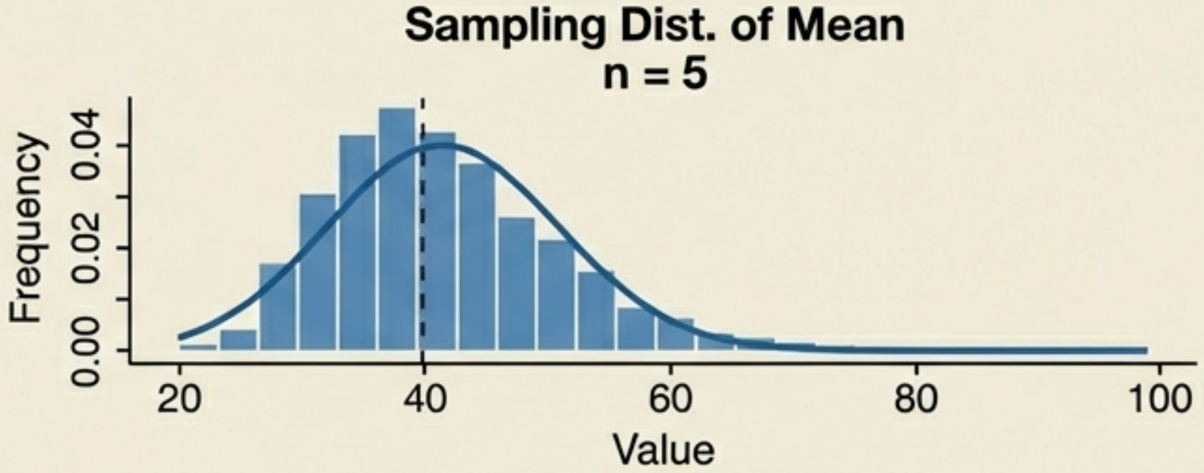


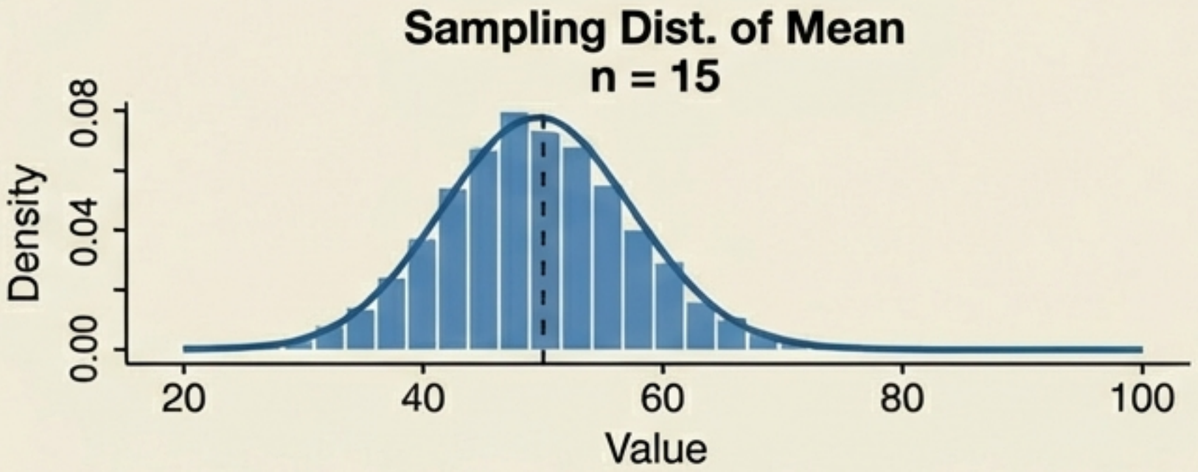
NotebookLM

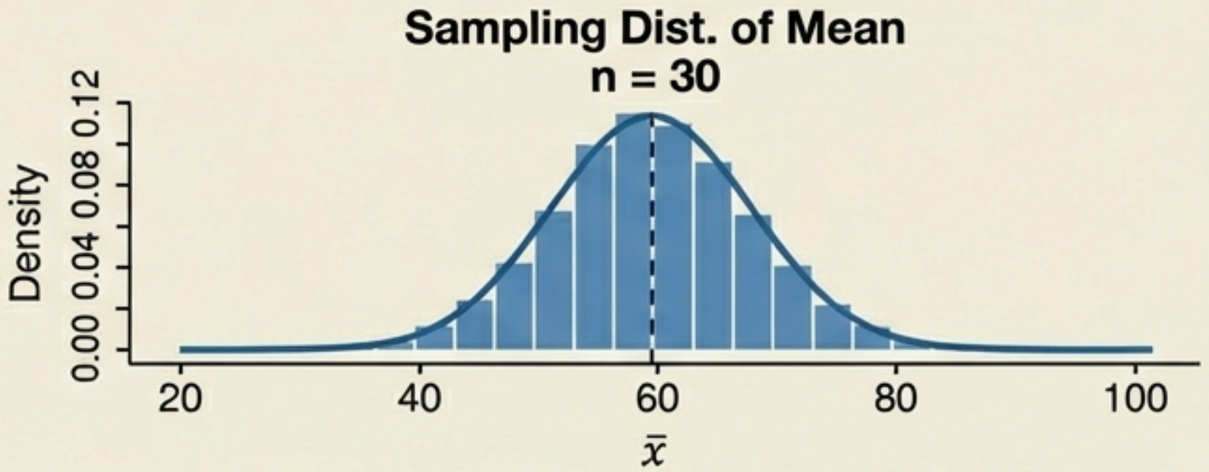
The CLT in Action: Order from Chaos

Observe a heavily right-skewed population. With a sample size of 5, the means are still skewed. But with a sample size of 30, the distribution of means transforms into a perfect bell curve.









NotebookLM

Quantifying Uncertainty: The Standard Error (SE)

We need a number to measure how precise our sample mean is. This number is the **Standard Error (SE)**. It represents the standard deviation of the sampling distribution.

- **Small SE**: Hypothetical samples are very similar = High Precision.

- **Large SE**: Hypothetical samples vary wildly = Low Precision.

Sample Standard Deviation

$$SE = \frac{s}{\sqrt{n}}$$

Square root of Sample Size

NotebookLM

Crucial Distinction: SD vs. SE

Standard Deviation (SD)

- Describes the spread of **individual data points**.
- Question: How different are the athletes from each other?
- Use: When describing the sample data.

Standard Error (SE)

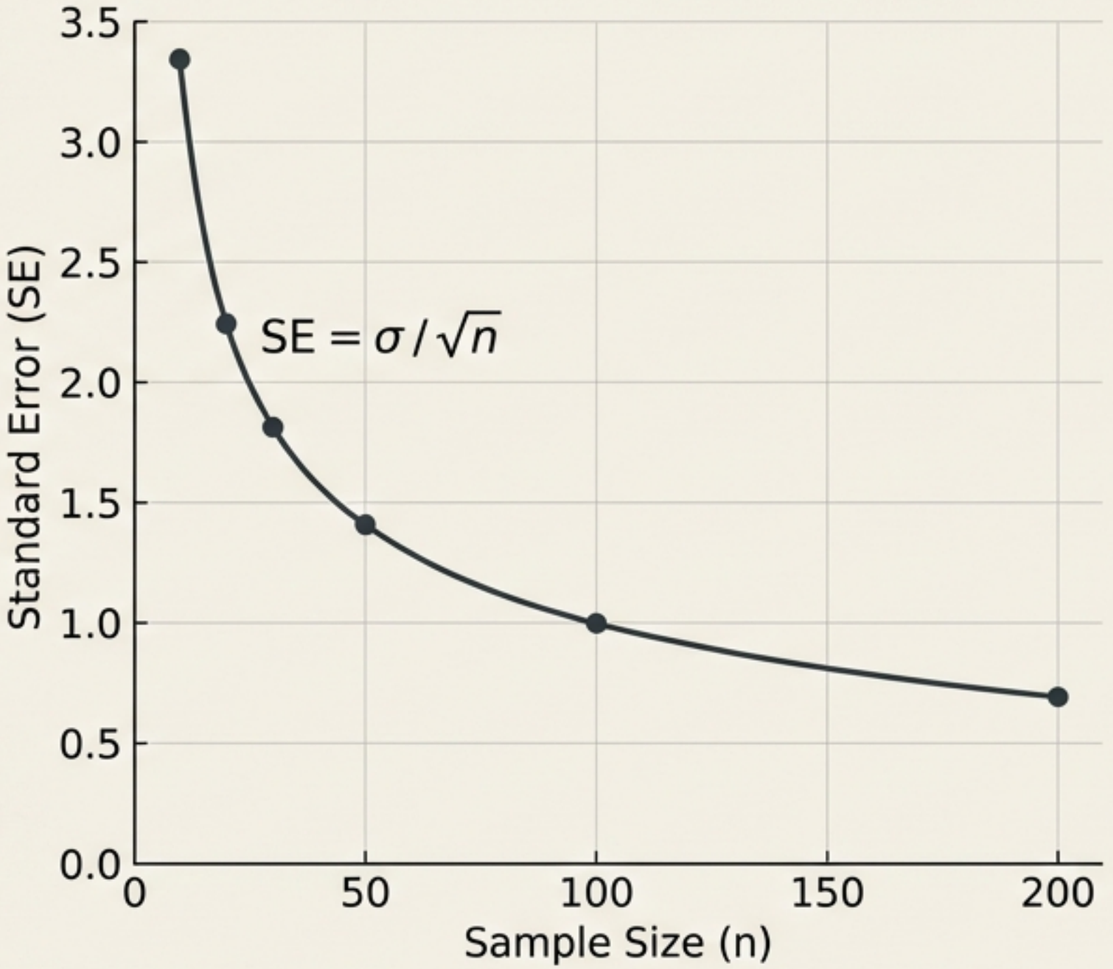
- Describes the precision of the **estimate**.
- Question: How far is our sample mean likely to be from the true population mean?
- Use: When making inferences about the population.

Rule of Thumb: Report SD for Description. Report SE for Inference.

NotebookLM

The Power of Sample Size (Diminishing Returns)

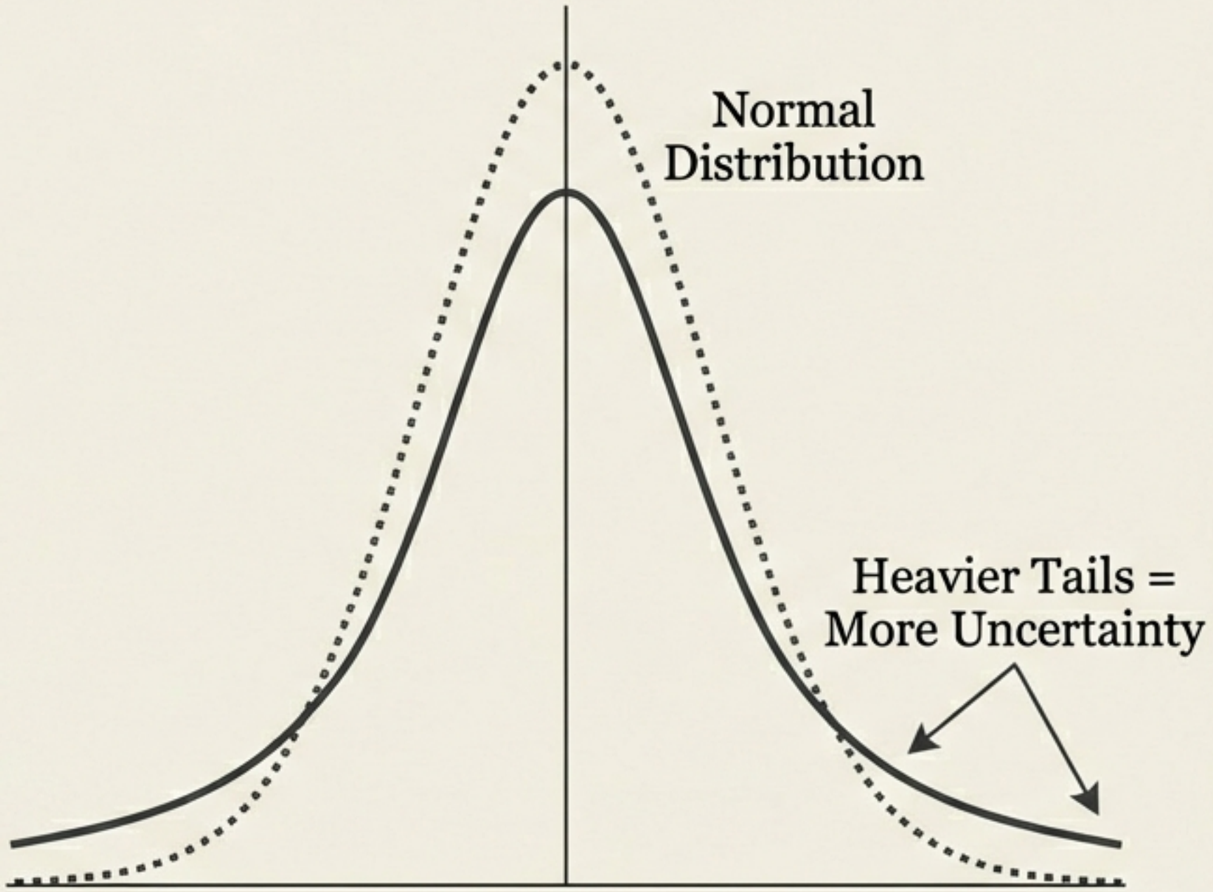
We can reduce uncertainty by increasing sample size, but it is not a linear relationship. The formula divides by the square root of n. To cut our error in half, we cannot just double our participants—we **must quadruple** them. This creates a law of diminishing returns where adding more participants eventually yields very little extra precision.



NotebookLM

Small Samples and the t-Distribution

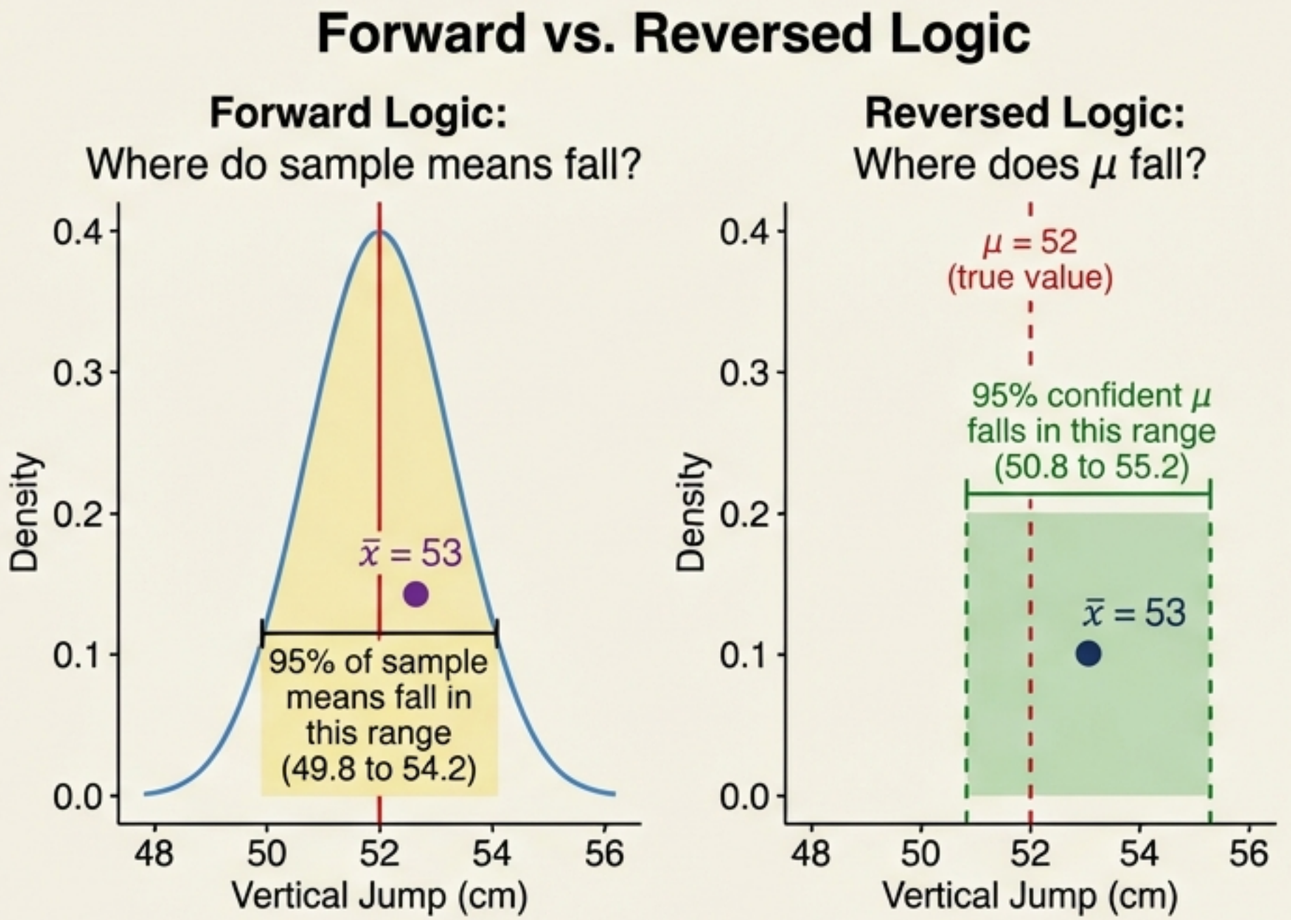
When sample sizes are small (typically $n < 30$), we cannot perfectly rely on the Normal Distribution because estimating the standard deviation introduces extra uncertainty. Instead, we use the **t-distribution**. It looks like a normal curve but has "heavier tails," acknowledging that with small samples, extreme errors are more likely.



NotebookLM

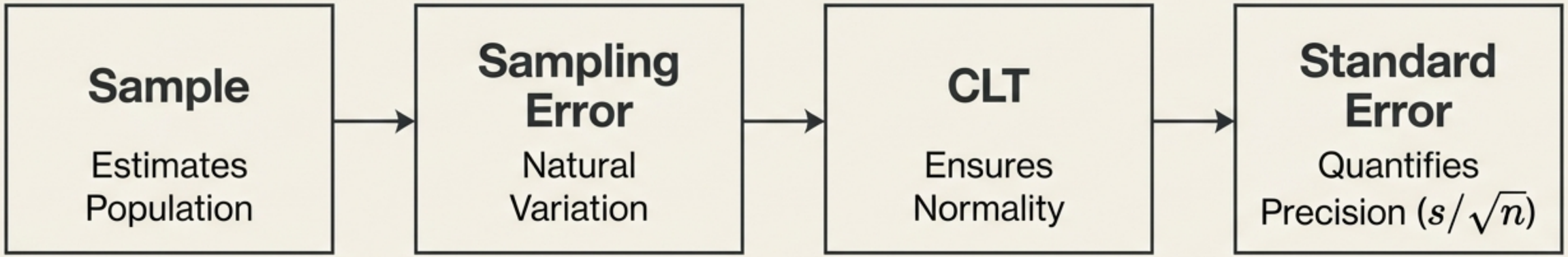
Closing the Loop: From Data to Inference

We use the sampling distribution to reverse our logic. If we know that 95% of all sample means fall within a certain distance of the true population mean, we assume that our specific sample mean is likely within that same distance.
This allows us to construct **Confidence Intervals.**



NotebookLM

Summary: The Bridge is Built



We have now built the bridge. In the following chapters, we will cross it to calculate specific Confidence Intervals (Chapter 9) and perform Hypothesis Testing (Chapter 10).

NotebookLM